Using the UMLS to Represent Medical Curriculum Content

Steven L. Kanter, M.D.
Office of Medical Education
University of Pittsburgh School of Medicine
Pittsburgh, Pennsylvania

ABSTRACT

Recent innovations in medical education have highlighted the need for faculty involved with the curriculum to carefully examine curricular content with goals of detecting omissions and unwanted redundancies of subject matter, adding and integrating new content, and deleting old content. A number of medical schools have attempted to deal with these issues by developing a database of curricular content information, most often using faculty- or student-selected keywords to represent each unit of instruction. However, several problems have been identified with this method, and achieving the goals mentioned above remains a formidable task. This paper outlines an alternative method that uses the resources of the UMLS to characterize a medical concept by the semantic types of its co-occurring terms. This approach can facilitate achievement of the aforementioned goals.

INTRODUCTION

A traditional medical school curriculum consists of a large amount of information presented by a large number of faculty. An ideal medical school curriculum is a dynamic entity by which students learn to access, manage, and utilize increasing amounts of rapidly changing information. In an attempt to move from a static traditional curriculum to a dynamic one, many schools have introduced new teaching formats, integrated courses with interdisciplinary faculty, centrally managed curricula, and other innovations. These innovations have highlighted the need for faculty involved with organizing, teaching, and managing the curriculum to carefully examine curricular content for purposes of detecting omissions and unwanted redundancies of subject matter, adding and integrating new content, and deleting old content [1,2,3,4,5].

Several medical schools have attempted to deal with these issues by developing a database of curricular content information [1,2,3,5,6,7,8,9]. Most groups defined a unit of instruction (e.g., a single lecture or a single laboratory session) and devised a system for representing the knowledge in that unit by selection of keywords or preparation of a summary by faculty and/or students. Keywords or other material were entered into a database management system, text file management system, or both. The MeSH vocabulary was often used for keyword selection, and at least one group was anticipating use of the UMLS (Unified Medical Language System) [3].

Several problems have been identified with the approach described above. Text file searches have the usual problems with precision and recall. Furthermore, Mattern et. al. [3] noted the need to "capture content with greater detail and richness." A few keywords per lecture is not adequate to represent the knowledge in an instructional unit. Adding to the problem is the lack of a controlled vocabulary that can capture the concepts expressed in medical curricular materials. MeSH was designed for indexing the biomedical literature and is not well-suited for educational purposes.

In order to successfully meet the goals of detecting omissions and unwanted redundancies of subject matter, adding and integrating new content, and deleting old content, the following must be addressed: (1) identification of knowledge in the instructional unit, (2) representation of the identified knowledge, (3) retrieval of curricular content information with acceptable precision and recall, and (4) the ability to make comparisons of curricular content to other information sources and databases.

For this paper, "(1)" is accomplished by the assumption that an instructional unit is a lecture, and the knowledge in the lecture is identified by a faculty-prepared lecture outline. To deal with "(2)", the resources of the UMLS are used to represent medical concepts in the instructional unit. Each medical concept is characterized by the semantic

0195-4210/92/$5.00

types of its co-occurring terms indexed in the medical literature during a specified period of time. This system of characterizing concepts enables clustering of similar concepts, and distinguishing dissimilar clusters of concepts, which allows comparisons of curricular content to other information sources ("(4)" above). Number "(3)" is not addressed in this paper.

METHODS

Two unrelated lectures were arbitrarily chosen from the first year curriculum at the University of Pittsburgh School of Medicine. One lecture was about "moods and emotions" given in a course called Behavioral Medicine. The other lecture was about the citric acid cycle given in a course entitled Cell Structure and Metabolism. A set of terms was arbitrarily selected from each lecture outline (six from "moods and emotions" -- namely, behavioral medicine, classification, emotions, health, quality of life, stress -- and seven from "citric acid cycle" -- namely acetyl coenzyme A, allosteric regulation, biochemistry, carbohydrates, carbon dioxide, catalysis, citric acid cycle) that represented concepts relevant to the lecture topic, and that were also MeSH main headings represented in the UMLS Metathesaurus 1.2 [10]. For each Meta-1.2 main concept that matched one of the terms chosen from a lecture outline, the following information was collected: (1) total number of articles indexed in MEDLINE (from January 1989 to April 1992) with the concept as a main heading (2) number of co-occurring terms classified by a semantic type that falls in one of eight top-level categories: anatomy, behaviors, chemicals and drugs, disease and pathologic processes, molecular biology, organisms, physiology, procedures [10, Appendix D].

Each concept was placed in a high-dimensional space, according to the numbers of co-occurring terms in each of the top-level semantic type categories. Euclidean distances were calculated between each pair of terms and placed in a proximity matrix. The proximity matrix is a table with a row and a column for each medical concept, so that every concept is compared to every other concept. In other words, the distance between every pair of concepts is recorded in the table. (It should be noted that distances were calculated using normalized data - i.e., the number of co-occurring terms in a particular semantic type category was divided by the total number of articles in which the concept appeared as a main heading.) The proximity data were then analyzed on a mainframe computer by a hierarchical cluster analysis (HCA) using average linkage (Statistical Package for the Social Sciences, Release 4.1). The cluster algorithm accepts proximity data (i.e., distances between medical concepts) and uses an agglomerative procedure to form progressively larger clusters of concepts.

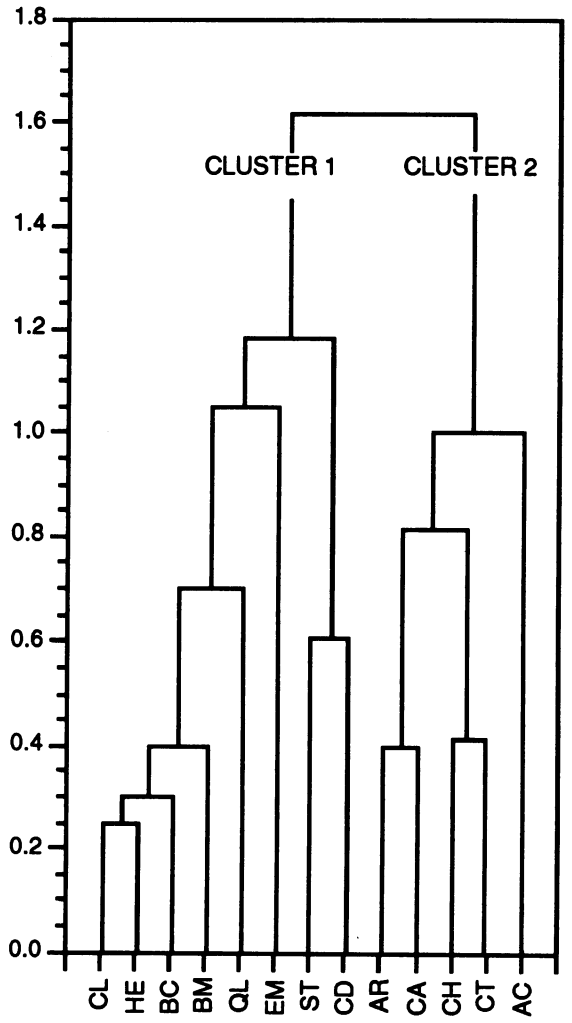

Figure 1: Dendrogram from hierarchical cluster analysis of medical concepts from two lectures. Abbreviations: CL = classification; HE = health; BC = biochemistry; BM = behavioral medicine; QL = quality of life; EM = emotions; ST = stress; CD = carbon dioxide; AR = allosteric regulation; CA = catalysis; CH = carbohydrates; CT = citric acid cycle; AC = acetyl coenzyme A.

RESULTS

Figure 1 shows a dendrogram from hierarchical cluster analysis using average linkage. Two large clusters are easily identified visually and are labelled CLUSTER 1 and CLUSTER 2. The number of clusters was verified by graphing fusion coefficient versus number of clusters and noting a "flattening in the curve" at the two-cluster solution [11].

The terms from each lecture (with two exceptions noted below) formed identifiable separate clusters. CLUSTER 1 contains all six terms from the "moods and emotions" lecture, but also contains two terms ("biochemistry" and "carbon dioxide") from the "citric acid cycle" lecture. CLUSTER 2 contains five of the seven terms from the "citric acid cycle" lecture. Although the term "biochemistry" did not cluster with other terms in its lecture, it did cluster with other "broad subject matter" type terms (i.e. behavioral medicine, health).

DISCUSSION

This paper presents an automated method for representing concepts from medical school lectures utilizing the resources of the UMLS [10]. A medical curricular concept is characterized by the semantic types of its co-occurring terms indexed in the medical literature during a specified period of time. The capability of clustering similar concepts, and distinguishing dissimilar clusters of concepts, provides a tool for analyzing curricular content.

Characterization of medical curricular concepts by the semantic types of their co-occurring terms allows comparison of concepts in the medical curriculum to indexed terms in the biomedical literature. This ability to explore the relationship between information in the medical curriculum and information in the biomedical literature is a prerequisite to the automated detection of omissions of subject matter. For example, if the term citric acid cycle is located in a high-dimensional space, all index terms from a given MEDLINE file can also be located in the space, and the pair-wise proximities can be subjected to hierarchical cluster analysis. The terms and clusters within a specified distance of the citric acid cycle cluster can then be examined, and a lecturer could make a decision about adding a new concept not previously considered. This data could also suggest areas of integration of material by noting concepts that cluster together.

There are several advantages to this approach compared to a database of keywords selected by faculty and/or students. The rapid discovery of new knowledge in biomedicine necessitates frequent updates to the content of a medical curriculum, making maintenance of a "keyword" database difficult, expensive, and time-consuming. The use of faculty content experts to select keywords significantly adds to the cost. In addition, the use of keywords is fraught with the usual problems of inadequate representation of content, as well as variability in descriptive term usage. In fact, Furnas, et. al. found that two people choose the same main keyword for a single well-known object less than 20% of the time [12]. Furthermore, in response to a query, a "keyword" database can only produce words with similar spellings, while a knowledge representation of clustered concepts would enable retrieval of a group of terms with similar "profiles", based on their location in clusters in high dimensional space. This could be extremely useful to a faculty member planning a lecture (or other lesson) who wants to know the content in which a topic was previously taught.

This project demonstrates how the resources of the UMLS can be utilized to explore the relationship between information in the medical curriculum and information in biomedicine (using the biomedical literature as a model of information in biomedicine). Based on this, one could envision an automated system to assist faculty in planning a lecture (or other learning session), an entire course, or offer decision support to select information to include in, and delete from, the curriculum. However, there are several issues requiring further inquiry. For one, an automated method to identify Meta-1.2 concepts in curricular text, such as the serial sliding-frame methodology used in R. Miller's CHARTLINE project [13], is needed. In addition, the characterization of a medical concept should be examined with different combinations of semantic type categories to determine the best combinations for use with large numbers of curricular terms. In other words, it will be important to construct a high-dimensional space that is "large enough" to accommodate all material from the pre-clinical curriculum and still have distinguishable clusters. Finally, it will be necessary to identify concepts important to medical education that are not well-

represented in the current version of MeSH. Because MeSH was designed for indexing the biomedical literature, it is not well-suited to educational purposes. However, Cooper's technique of relating text phrases to MeSH indexing terms in a probabilistic manner may be one method of depicting non-represented terms, thus allowing better representation of the knowledge in a medical school curriculum.

Reference

1. Curry L, et al. Computerization of Undergraduate Medical Curriculum Content. *Medical Education*. 18:71-74, 1984.

2. Gotlib D, DesGroseilliers J, Dolphin P. A Computerized Database-Management System for Curriculum Analysis. *Canadian Medical Association Journal*. 131:861-863, 1984.

3. Mattern WD, et al. Computer Databases of Medical School Curricula. *Academic Medicine*. 67 (1):12-16, January 1992.

4. Piggins JL, et al. A Computer-Based System for Indexing Curriculum Content. *Proceedings of the Fourteenth Annual Symposium on Computer Applications in Medical Care*. R.A. Miller (ed.) IEEE Computer Society Press, 1990; 210-214.

5. Rozinski EF, Blanton WB. A System of Cataloging the Subject Matter Content of a Medical School Curriculum. *Journal of Medical Education*. 37:1092-1100, 1962.

6. Buckenham S, Sellers EM, Rothman AI. An application of Computers To Curriculum Review and Planning. *Journal of Medical Education*. 61:41-45, 1986.

7. Mann D, et al. A curriculum Database with Boolean Natural-Language Searching in Hypercard. *Proceeding of the Sixteenth Annual Symposium on Computer applications in Medical Care*. Baltimore, MD. M.E. Frisse, M.D. (Ed.) November 1992.

8. Mattern WD, et al. A Computerized Representation of a Medical School Curriculum: Integration of Relations and Text Management Software in Database Design. *Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care*. P.D. Clayton (ed.) McGraw Hill, 1992; 323-327.

9. Rosen RL, et al. Using a Database to Analyze Core Basic Science Content in a Problem-based Curriculum. *Academic Medicine*. 67 (8): 535-538, August 1992.

10. *Unified Medical Language System*. 3rd experimental edition, 1992. National Library of Medicine, Bethesda, Maryland.

11. Aldenderfer MS, Blashfield RK. *Cluster Analysis*. Sage University Paper on Quantitative Applications in the Social Sciences, 07-44. Beverly Hills, California, 1984.

12. Furnas GW, et al. The Vocabulary Problem in Human-System Communications. *Communications of the ACM*. 30:964-971, 1987.

13. Miller RA, et al. ©HARTLINE: Providing Bibliographic References Relevant to Patient Charts Using the UMLS Metathesaurus Knowledge Sources. *Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care*. Baltimore, MD. M.E. Frisse, M.D. (Ed.) November 1992.